frontiers

Machine learning approach for classifying Multiple Sclerosis courses by combining clinical data with lesion loads and Magnetic Resonance metabolic features

Adrian Ion-Mărgineanu [1,2,3,*], **Gabriel Kocevar**[1], **Claudio Stamile** [1,2,3], **Diana M Sima**[2,3,4], **Françoise Durand-Dubief**[1,5], **Sabine Van Huffel**[2,3], **and Dominique Sappey-Marinier**[1,6]

[1] *CREATIS CNRS UMR5220 & INSERM U1206; Université de Lyon, Université Claude Bernard-Lyon 1, INSA-Lyon, Villeurbanne, France*
[2] *KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium*
[3] *imec, Leuven, Belgium*
[4] *ico**metrix**, R&D department, Leuven, Belgium*
[5] *Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France*
[6] *CERMEP - Imagerie du Vivant, Université de Lyon, Bron, France*

Correspondence*:
Adrian Ion-Mărgineanu, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
adrian@esat.kuleuven.be

ABSTRACT

Purpose. The purpose of this study is classifying multiple sclerosis (MS) patients in the four clinical forms as defined by the McDonald criteria using machine learning algorithms trained on clinical data combined with lesion loads and magnetic resonance metabolic features.

Materials and Methods. Eighty-seven MS patients (12 Clinically Isolated Syndrome (CIS), 30 Relapse Remitting (RR), 17 Primary Progressive (PP) and 28 Secondary Progressive (SP)) and eighteen healthy controls were included in this study. Longitudinal data available for each MS patient included clinical (e.g. age, disease duration, Expanded Disability Status Scale), conventional magnetic resonance imaging and spectroscopic imaging. We extract *N*-acetyl-aspartate (NAA), Choline (Cho), and Creatine (Cre) concentrations, and we compute three features for each spectroscopic grid by averaging metabolite ratios (NAA/Cho, NAA/Cre, Cho/Cre) over good quality voxels. We built linear mixed-effects models to test for statistically significant differences between MS forms. We test nine binary classification tasks on clinical data, lesion loads, and

metabolic features, using a leave-one-patient-out cross-validation method based on 100 random patient-based bootstrap selections. We compute F1-scores and BAR values after tuning Linear Discriminant Analysis (LDA), Support Vector Machines with gaussian kernel (SVM-rbf), and Random Forests.

Results. Statistically significant differences were found between the disease starting points of each MS form using four different response variables: Lesion Load, NAA/Cre, NAA/Cho, and Cho/Cre ratios. Training SVM-rbf on clinical and lesion loads yields F1-scores of 71-72% for CIS vs. RR and CIS vs. RR+SP, respectively. For RR vs. PP we obtained good classification results (maximum F1-score of 85%) after training LDA on clinical and metabolic features, while for RR vs. SP we obtained slightly higher classification results (maximum F1-score of 87%) after training LDA and SVM-rbf on clinical, lesion loads and metabolic features.

Conclusions. Our results suggest that metabolic features are better at differentiating between relapsing-remitting and primary progressive forms, while lesion loads are better at differentiating between relapsing-remitting and secondary progressive forms. Therefore, combining clinical data with magnetic resonance lesion loads and metabolic features can improve the discrimination between relapsing-remitting and progressive forms.

Keywords: multiple sclerosis, longitudinal analysis, magnetic resonance spectroscopic imaging, EDSS, lesion load, machine learning

1 INTRODUCTION

Multiple sclerosis (MS) is an inflammatory disorder of the brain and spinal cord in which focal lymphocytic infiltration leads to damage of myelin and axonsCompston and Coles (2008). MS affects approximately 2.5 million people worldwide, with an onset age commonly between 20 and 40 years, and an incidence more than twice as high in women compared to menMcAlpine and Compston (2005).

The majority of MS patients (85%) usually experience a first attack defined as Clinically Isolated Syndrome (CIS), and will develop a relapsing-remitting (RR) formMiller et al. (2012). Two thirds of the RR patients will develop a secondary progressive (SP) form, while the other third will follow a benign courseScalfari et al. (2010). The rest of MS patients (15%) will start directly with a primary progressive (PP) form.

The criteria to diagnose MS forms was originally formulated by McDonald in 2001McDonald et al. (2001) and revised by Polman in 2005Polman et al. (2005) and 2011Polman et al. (2011). They all rely on using conventional magnetic resonance imaging techniques (MRI) such as T1-weighted, gadolinium-enhanced T1-weighted MRI, as well as T2-weighted and FLAIR, due to a high sensitivity for visualizing MS lesions. Conventional MRI is also used for quantifying lesion load (LL), a marker of inflammation process but only a moderate predictor of MS evolutionFilippi et al. (1994).

More recently, advanced magnetic resonance techniques such as ^{1}H-Magnetic Resonance Spectroscopic Imaging (MRSI), Diffusion Tensor Imaging (DTI) and Magnetization Transfer Imaging (MTI) have been shown Rovira et al. (2013) to provide a better characterization of the normal appearing white matter (NAWM) and thus a better understanding of the pathological mechanisms of MS. MTI metrics reflect the demyelination and remyelination processes and have been shown to predict the evolution of MS lesions. DTI metrics are very sensitive to the MS pathology and have been shown to be mainly affected by myelin loss and decreased neuronal integrity. MRS metrics provide high MS pathological specificity as well as high sensitivity to biochemical changes. Decrease of *N*-acetyl-aspartate (NAA) was observed in both

chronic lesions and NAWM, reflecting a neuronal integrity lossRovira et al. (2013). Choline (Cho) and Creatine (Cre) contents were found to be increased in WM lesions and in NAWM, indicating the presence of severe demyelination and cell proliferation in relation with inflammatory processesTartaglia et al. (2002); Sajja et al. (2009).

Therefore, in this study we investigate the added value of magnetic resonance metabolic features (NAA/Cho, NAA/Cre, Cho/Cre) combined with routinely collected clinical MS data (e.g. patient age, disease duration (DD), Expanded Disability Status Scale (EDSS)) and lesion load values (LL). To this purpose, we build multiple binary classifiers to automatically discriminate between different clinical forms of MS patients, by training each classifier on combinations of clinical data, lesion loads and metabolic features.

2 MATERIALS AND METHODS

2.1 Patient population

Eighty-seven MS patients (12 CIS, 30 RR, 28 SP and 17 PP) were included in this study, while 18 volunteers without any neurological disorders served as healthy control (HC) subjects. Diagnosis and disease course were established according to the McDonald criteria Lublin et al. (1996); McDonald et al. (2001). This prospective study was approved by the local ethics committee (CPP Sud-Est IV) and the French national agency for medicine and health products safety (ANSM) and written informed consents were obtained from all patients and control subjects prior to study initiation. More details for each MS group, such as average age at first scan, average disease duration, median EDSS and average lesion loads can be found in Table 1.

	CIS	RR	PP	SP
Number of patients (Male/Female)	12 (6/6)	30 (6/24)	17 (6/11)	28 (17/11)
Age at first scan [years]	31.8 (6.4)	33.2 (7)	39.5 (6)	41.1 (4.8)
Disease duration [years]	2.9 (1.9)	8.3 (4.8)	7.5 (2.9)	14.9 (6.1)
EDSS median [range]	1 (0-4)	2 (0-5.5)	4 (2-7.5)	5 (3-8.5)
Lesion Load [ml]	6.6 (3.5)	16.7 (12.6)	20.8 (13)	31 (12.9)
Total number of scans	62	226	125	206

Table 1. Patient population: Age - average value (standard deviation); Disease duration - average value (standard deviation); EDSS - median (minimum - maximum); Lesion Load - average value (standard deviation).

2.2 Longitudinal MS data

The MS patients involved in this study were scanned multiple times over a different period for each patient, ranging from 2.5 to 6 years. The minimum number of scans is 3, while the maximum is 10. The gap between two consecutive scans is either 6 months or 1 year. In total there are 619 MS scans, but because of missing lesion loads and metabolic features, there are 592 (95.6%) scans with full complete data, leading to an average of 6-7 complete scans/patient.

2.3 MRI acquisition and processing

All patients and control subjects underwent MR examination using a 1.5 Tesla MR system (Sonata Siemens, Erlangen, Germany) and an 8 elements phased-array head-coil.

2.3.1 Conventional MRI

Conventional MRI protocol consisted of a 3 dimensional T1-weighted (magnetization prepared rapid gradient echo-MPRAGE) sequence with repetition time/echo time/time for inversion (TR/TE/TI)= $1970/3.93/1100$ ms, flip angle= $15°$, matrix size= 256×256, field of view (FOV)= 256×256 mm, slice thickness= 1 mm, voxel size= $1 \times 1 \times 1$ mm, acquisition time= 4.62 min, and a fluid attenuated inversion recovery (FLAIR) sequence with TR/TE/TI= $8000/105/2200$ ms, flip angle= $150°$, matrix size= 192×256, field of view (FOV)= 240×240 mm, slice thickness= 3 mm, voxel size= $0.9 \times 0.9 \times 3$ mm, acquisition time= 4.57 min.

2.3.2 MRSI acquisition

MRSI data was acquired from one slice of 1.5 cm thickness, placed above the corpus callosum and along the anterior commissure - posterior commissure (AC-PC) axis, encompassing the centrum semiovale region, and took 5 minutes and 20 seconds. A point-resolved spectroscopic sequence (PRESS) with TR=1690 ms and TE=135 ms was used to select a volume of interest (VOI) of $105 \times 105 \times 15$ mm^3 during the acquisition of 24×24 (interpolated to 32×32) phase-encodings over a field of view (FOV) of 240×240 mm^2.

2.3.3 MRSI processing

MRSI data processing was performed using SPID Poullet (2008); Poullet et al. (2008) in MatLab 2015a (MathWorks, Natick, MA, USA). AQSES-MRSI Poullet et al. (2007); Sava et al. (2011) was used to quantify *N*-acetyl-aspartate, Choline (Cho), and Creatine (Cre), using a synthetic basis set. The basis set incorporates prior knowledge of the individual metabolites in the quantification procedure. MPFIR (maximum-phase finite impulse response) filtering Sundin et al. (1999) was included in the AQSES-MRSI procedure for residual water suppression, with a filter length of 50 and spectral range from 1.9 to 3.4 ppm. A band of two voxels at the outer edges of each VOI was discarded in order to avoid chemical shift displacement artifacts and lipid contamination artifacts.

2.3.4 Quality control

After quantifying metabolites from all MRSI grids, a quality control was performed. Voxels with Cramer-Rao Lower Bounds (CRLBs) lower than 10% for *each* of the three metabolites (NAA, Cho, and Cre) were kept as having "good quality" to perform feature extraction. If the number of "good quality" voxels is lower than 50% of the total amount of voxels in the MRSI grid, then the acquisition is discarded. All 18 Control subjects had MRSI data with a number of "good quality" voxels higher than 50% of the total amount of voxels, and 606 out of 619 (97.9%) MRSI data from MS patients had good quality as defined earlier.

2.4 Feature extraction

In this study we use three types of features: clinical (e.g. patient age, disease duration, and EDSS), lesion loads, and metabolic features. The clinical features are routinely acquired in the hospital. The lesion loads were computed based on T1 and FLAIR, using the MSmetrix software Jain et al. (2015) developed by ico**metrix** (Leuven, Belgium). The computation of metabolic features was performed in two steps: three metabolic ratios (NAA/Cho, NAA/Cre, Cho/Cre) were computed for each "good quality" voxel and then averaged, leading to three metabolic features extracted from each MRSI grid.

2.5 Training approach

Nine binary classification tasks were studied: HC vs. CIS, HC vs. RR, HC vs. PP, HC vs. RR+SP, HC vs. PP+SP, CIS vs. RR, CIS vs. RR+SP, RR vs. PP, RR vs. SP. The first three tasks investigated differences between HC and the starting MS forms (CIS, RR, and PP). The next task investigated differences between HC and MS patients that are likely to evolve or had evolved into secondary progressive form (RR+SP). Afterwards, we investigated differences between HC and definite progressive forms (PP+SP). The next two tasks investigated differences between CIS patients and the most likely progression of CIS, namely RR and RR+SP. From a neurological point of view, the last two tasks were the most intriguing, as they were discriminating between the most common inflammatory MS form (RR) and the two progressive forms, PP and SP.

For each task, data normalization was performed. We use a leave-one-patient-out cross-validation (LOPOCV) setup combined with 100 random patient-based bootstrap selections for the training set. In this way, the test set has all data points of one patient, while the training set has $n-1$ data points corresponding to $n-1$ patients, where n is the total number of patients, different for each classification task (e.g. for HC vs. CIS, $n=30$). Basically, to construct the training set, we randomly select one data point from each patient assigned to the training set. The test set always includes all data points of the test patient. We repeat the procedure 100 times and store the results. Each data point from the test set will be assigned 100 times to either class 1 or class 2, and in the end it will be assigned to one of the classes according to majority voting. This procedure is repeated until all patients from each classification task have been tested.

By using this random patient-based bootstrap selection, the two classes in the training set have a more balanced distribution of points (18 HC, 12 CIS, 30 RR, 17 PP, 28 SP), compared to using the total number of points of each class (18 HC, 61 CIS, 214 RR, 121 PP, 196 SP).

2.6 Performance measures and statistical testing

For each task, we computed and reported four measures, in percentage: F1-score, sensitivity, specificity, and balanced accuracy rate (BAR). We explain these four measures using the general confusion matrix in Table 2.

Confusion matrix		predicted condition	
		predicted negative	predicted positive
true condition	condition negative	true negative (TN)	false positive (FP)
	condition positive	false negative (FN)	true positive (TP)

Table 2. General confusion matrix.

The four measures are defined by the following formulas: $F1 = \frac{2\times TP}{2\times TP+FN+FP}$, $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, $BAR = \frac{Sensitivity+Specificity}{2}$.

Throughout our study the positive class was the first class from each of the nine binary classification tasks: HC for the first 5 tasks, CIS for the 6^{th} and 7^{th} tasks, and RR for 8^{th} and 9^{th} tasks.

In order to correctly assess if there are significant differences between the four MS groups, we built several linear mixed effects models which were able to incorporate the temporal evolution of each patient's MS course. We used five fixed effects and two random effects. The fixed effects are: MS course, gender, disease onset age, disease duration, and the interaction between MS course and disease duration. The

random effects are set for each patient allowing an individual starting point and an individual disease evolution. The most interesting fixed effect for this study is the first one, which represents the average of the response variable at the beginning of the MS course, or when 'disease duration' $= 0$. We built four linear mixed effects models, one for each response variable: NAA/Cho, NAA/Cre, Cho/Cre, and lesion load. All statistical models were built in the 'R' software using the "lme4" package Bates (2010), statistical testing was done using the "lmerTest" package Kuznetsova et al. (2015) and post-hoc analysis was done using the "multcomp" package Hothorn et al. (2008). All tests were done for a significance level (α) of 0.05.

2.7 Classifiers

Three supervised classifiers implemented in Python 2.7.11 with scikit-learn 0.17.1 Pedregosa et al. (2011) have been used: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Random Forest (RF). We tuned each classifier's parameters by optimizing the F1-score over a 5-fold cross validation on the training set within a grid search of individual parameters, specified further for each particular classifier. Fisher's LDA Fisher (1936) is based on a linear combination of input features, with three possible solvers: singular value decomposition, least squares solution, and eigenvalue decomposition. Tuning involved choosing between the first solver and the last two solvers combined with shrinkage varying from 0 to 1 in steps of 0.1. Class unbalance was adjusted by setting the *priors* parameter equal to class probabilities. We use SVM Cortes and Vapnik (1995); Cristianini and Shawe-Taylor (2000) with a radial basis function kernel (SVM-rbf), defined by two parameters: C, or the misclassification cost, and γ, which is proportional to the inverse of a support vector's radius of influence. We tuned C and γ by performing a logarithmic grid search between 0.00001 and 100000. Class unbalance was adjusted by setting the *class_weight* parameter to *balanced*. Random Forests Breiman (2001) is based on a group of decision trees. We tune the number of decision trees between 200, 400, 600, 800, and 1000. Class unbalance was adjusted by setting the *class_weight* parameter to *balanced_subsample*.

3 RESULTS

Figure 1 shows boxplots comparing MR metabolic features (A, B, C) and lesion loads (D) extracted from HC and each MS course. Boxplots are drawn using default style in MatLab, meaning the middle line inside the box represents the median value, the vertical limits are the 25^{th} and 75^{th} percentiles (q_1 and q_3), each whisker covers 1.5 the interquartile range (i.e. $q_3 - q_1$), and the crosses outside the whiskers represent outliers. Figures 2, 3, 4, and 5 from Appendix show the MS data points in various 2-D feature spaces.

Using the previously described (Section 2.6) linear mixed-effects models we found that the fixed effect MS course is statistically significant in the evolution of NAA/Cho, NAA/Cre, Cho/Cre, and LL, with corresponding p-values of: 3.4×10^{-6}, 2×10^{-4}, 2.3×10^{-2}, and 2.6×10^{-4}. Table 3 provides adjusted p-values for multiple comparisons between the MS groups.

Table 4 shows F1-scores after training LDA using only metabolic ratios, as clinical data and lesion loads were not available for healthy controls. Corresponding BAR, sensitivity and specificity values of this table can be found in Table 6 in Appendix. If F1-scores are missing, then the classifier assigned all data points to the negative class (second MS group).

Surprisingly, the F1-scores for separating HC from any MS course are very low, and the same holds true for separating very early MS form (CIS) and the most likely MS evolution, RR and RR+SP. In contrast,

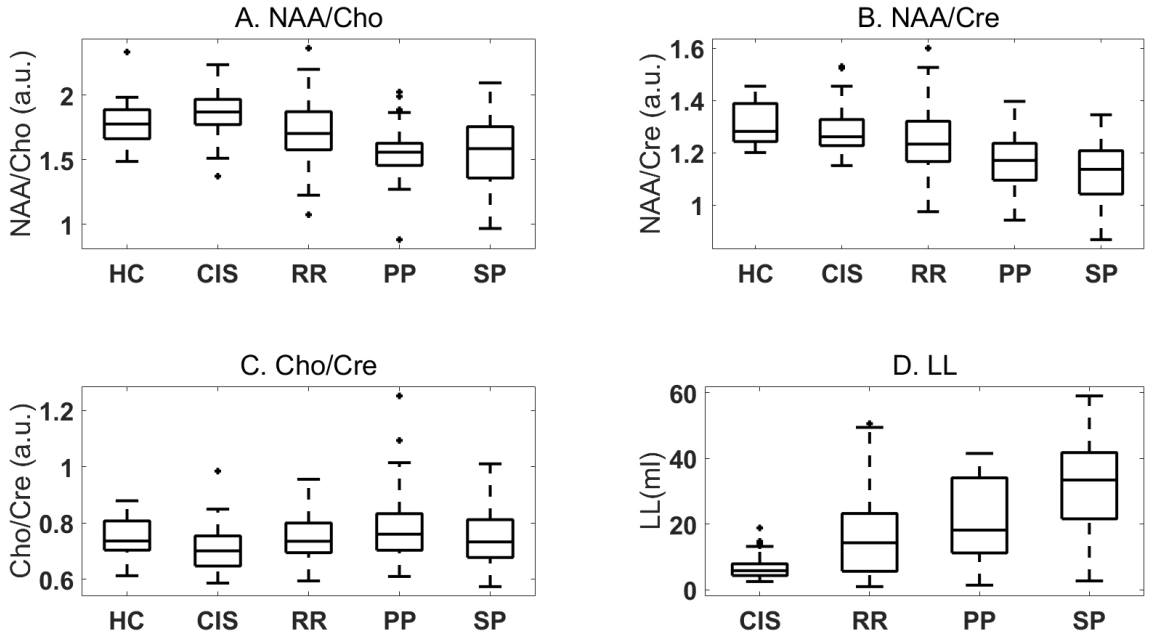

Figure 1. Boxplots of MR metabolic features and lesion loads extracted from HC and MS patients: A. NAA/Cho; B. NAA/Cre; C. Cho/Cre; D. Lesion load (LL).

	CIS - RR	RR - PP	RR - SP
NAA/Cho	-	**	**
NAA/Cre	-	-	*
Cho/Cre	-	-	-
LL	-	-	*

Table 3. Adjusted p-values for multiple comparisons between MS groups modelled by linear mixed effects model, tested using the "multcomp" package in 'R' (* for $p < 0.05$ and ** for $p < 0.01$).

	NAA/Cho	NAA/Cre	Cho/Cre	All 3 metabolic ratios
HC vs. CIS	35	33	43	36
HC vs. RR	6	16	-	14
HC vs. PP	47	45	19	49
HC vs. RR+SP	8	19	-	16
HC vs. PP+SP	21	26	-	28
CIS vs. RR	15	-	-	21
CIS vs. RR+SP	3	-	-	19
RR vs. PP	75	78	75	74
RR vs. SP	60	67	58	69

Table 4. F1-scores for all nine classification tasks (rows) after training LDA using only metabolic ratios. Values above 75 are coloured in light gray.

for RR vs. PP we find that all three metabolic ratios have F1-scores higher than 75, with a maximum of 78 for NAA/Cre. For RR vs. SP the F1-scores are lower, with a maximum of 69 after combining all metabolic features.

Table 5 shows F1-scores of classification tasks involving only MS patients. Training was done on seven different combinations of features to evaluate the classification power of clinical data, lesion loads, and metabolic features. Corresponding BAR, sensitivity, and specificity values can be found in Appendix in

Tables 7, 8, and9, respectively. If F1-scores are missing, then the classifier assigned all data points to the negative class (second MS group).

	CIS vs. RR			CIS vs. RR+SP			RR vs. PP			RR vs. SP		
	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF
M	21	48	11	19	31	-	74	52	73	69	70	67
LL	-	51	27	-	40	24	71	19	73	75	77	68
Age + DD	48	58	51	44	56	50	79	64	74	76	75	71
Age + DD + EDSS	55	65	49	57	66	48	85	81	79	84	85	84
Age + DD + EDSS + LL	67	71	59	63	72	60	79	75	79	86	86	86
Age + DD + EDSS + M	56	59	48	60	59	51	85	83	80	86	87	85
Age + DD + EDSS + LL + M	65	64	57	65	63	57	83	81	78	87	87	86

Table 5. F1-scores for classification tasks involving only MS patients (columns). Abbreviations: M = all three average metabolic ratios; Age = patient age; DD = disease duration; LL = lesion load; EDSS = Expanded Disability Status Scale. Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, while values larger than 85 are coloured in dark gray.

The highest F1-scores for CIS vs. RR and CIS vs. RR+SP, respectively 71 and 72, were achieved by SVM-rbf trained on clinical data and lesion loads. Training any classifier only on metabolic features yielded very low F1-scores.

The highest F1-score for RR vs. PP (85) was achieved by LDA using patient age, disease age, and EDSS. Adding all spectroscopic information maintained the F1-score at 85, while adding lesion load lowered the F1-score at 79. LDA outperformed SVM-rbf and RF in all RR vs. PP cases, always achieving an F1-score higher than 70.

The highest value for RR vs. SP (87) was first achieved after training SVM-rbf on clinical and metabolic features, but also with LDA trained on all features combined (clinical data, lesion loads, and metabolic features). SVM-rbf outperfomed LDA in the majority RR vs. SP cases, but only with 1 to 2%.

4 DISCUSSION

In this paper, we present results for nine binary classification problems using clinical data, lesion loads and metabolic features extracted from MS patients and healthy controls. We focused on metabolic features as numerous studies showed significant metabolic alterations in MS patients of different MS forms. It has been demonstrated that metabolic abnormalities in MS patients are not restricted to lesions alone Narayana et al. (1998); Doyle et al. (1995); He et al. (2005); Fu et al. (1998); Husted et al. (1994); Narayanan et al. (1997); Sarchielli et al. (1999) and NAWM tissue is well known to be altered in MS Narayana (2005); De Stefano and Filippi (2007). Concentrations of NAA in NAWM were shown to be significantly lower in MS patients Bitsch et al. (1999); Bjartmar et al. (2001); Tiberio et al. (2006); Inglese et al. (2003); Suhy et al. (2000); Wattjes et al. (2007, 2008). Concentrations of Cho and Cre in NAWM were shown to be significantly higher in MS patients Narayana et al. (1998); Tartaglia et al. (2002); Inglese et al. (2003); Tourbah et al. (1999); Suhy et al. (2000). Concentrations of NAA/Cre in NAWM were shown to be significantly lower in MS patients Leary et al. (1999); Narayana et al. (2004). Multiple studies also report significant differences between metabolite concentrations in lesions vs. NAWM of HC: lower NAA and increased Cho and Cre Narayana et al. (1998); Davie et al. (1997); He et al. (2005); Arnold et al. (2000); Wolinsky et al. (1990); Larsson et al. (1991); Davie et al. (1994).

Our findings are in agreement with these previous reports as decreased NAA and increased Cho and Cre contents were measured in NAWM and lesions of MS patients. After building linear mixed-effects models

to properly analyze the statistical difference between the four clinical courses, we observed significant differences at the disease starting points of all MS courses using four response variables, namely the lesion load, NAA/Cre, NAA/Cho, and Cho/Cre ratios. A cross-sectional study Hannoun et al. (2012) based on a subset of our MRSI data found statistical differences in the NAA/Cre and NAA/Cho ratios between HC and RR, PP, SP, and RR+PP+SP patients. To our knowledge, there is only one study that reports sensitivity and specificity values for classifying healthy controls from MS patients based on spectroscopic features. Inglese et al. show in Inglese et al. (2003) that absolute values of choline in NAWM can differentiate 9 controls and 10 out of 11 RR patients.

Other MS classification studies are Muthuraman et al. (2016) and Kocevar et al. (2016), both based on diffusion features. The first one reports a classification accuracy of 97% between 20 CIS and 33 RR patients. The second one analyzes classification tasks based on DTI data from a cross-sectional subset of our database. They found very high F1-scores (91.8% for both HC-CIS and CIS-RR) after training SVM-rbf on six global brain connectivity metrics. For RR vs. PP their maximum F1-score was 75.6%, which is lower than our results based on metabolic features, while for RR vs. SP, their maximum F1-score was 85.5%, which is comparable to our results. It is also worth mentioning that they did not use any clinical data, which might improve their results.

In this study, we analyzed the added value of combining standard clinical data with quantitative magnetic resonance features. To this purpose, we trained linear and non-linear classifiers only on advanced MR features, and then only on clinical data. Afterwards we train the classifiers on clinical data combined with lesion loads and metabolic features.

Although MS patients are expected to have significantly different WM metabolism compared to healthy controls, this difference was not reflected in the metabolic average obtained from "good quality" voxels (Figure 2, A and B). This result is not entirely surprising, considering that we averaged over a high number of voxels, and the subtle lesion information could be lost in the average. However, we can visually see in Figure 2:C&D that the two progressive MS courses tend to have lower NAA/Cho and NAA/Cre ratios than healthy controls.

CIS and RR patients' distribution over the NAA/Cho and NAA/Cre feature space do not differ much, as seen in Figure 3:A. Disease duration interval for RR patients is much larger than for CIS patients, as most of CIS patients have a disease duration lower than 5 years, which can be seen in Figure 4:A. Because RR patients have more relapses than CIS patients, the number of lesions will be higher and the lesion volume as well, while EDSS scores are mainly in the same range, as seen in Figure 5:A. BAR values in Table 7 show a maximum of 85, when combining patient age, disease duration, EDSS, and lesion load. However, the corresponding maximum F1-score of 71 is much lower because the dataset is unbalanced (61 CIS vs. 214 RR), heavily influencing the classifier's precision. In this case the F1-score reflects better than BAR the difficulty of discriminating CIS from RR forms.

CIS and SP patients' distribution over different features is visible in Figure 3:B, Figure 4:B, and Figure 5:B, and it is clear that these two are the least and most advanced forms of MS. Because RR patients will eventually evolve into SP forms during their lifetime, we grouped together RR and SP patients for a separate classification task versus CIS patients. BAR values in Table 7 show a maximum of 92, when combining patient age, disease duration, EDSS and lesion load. The same discussion as for CIS vs. RR apply: the corresponding maximum F1-score is only 72 because the dataset is very unbalanced (61 CIS vs. 410 RR+SP) and the precision will be very low.

RR and PP patients can be discriminated using only EDSS by visually inspecting Figure 5:C. Training a linear classifier on clinical data (patient age, disease duration, and EDSS) gives the maximum F1-score of 85. Adding the 3 metabolic features keeps the score at 85, while adding lesion load information lowers the score to 79. This drop in the F1-score suggests that lesion load is not useful in differentiating RR from PP patients. Indeed, these two MS forms have the closest lesion load averages (16.7 ml and 20.8 ml), as shown in Table 1. In contrast, the clinical status of RR and PP patients are very different, as reflected by the EDSS values of 2 for RR and 4 for PP. Moreover, training LDA on individual metabolic features always provided higher F1-scores than lesion load, therefore we can conclude that for RR vs. PP, metabolic features have a higher discrimination power than LL. BAR values in Table 7 are also closer to the F1-scores in Table 5 because the dataset is more balanced compared to previous cases.

RR and SP patients can also be discriminated using only EDSS by visually inspecting Figure 5:D. Our results showed that EDSS is very important in differentiating RR patients from primary or secondary progressive patients. We also report consistent higher F1-scores for classifiers trained only on lesion load compared to classifiers trained only on metabolic features. Furthermore, it is clearly visible in Table 4 that we obtain higher F1-scores for this classification task using multiple features, compared to the rest of 8 tasks. These findings suggest that in the future it might be possible to build a decision support system using clinical data combined with lesion loads and metabolic features.

However, this study suffers from a few limitations, one of them being the low scanning frequency of only 1.5 Tesla. Firstly, it is known that the sensitivity of lesion load segmentation is improved by scanning at higher frequenciesSicotte et al. (2003). Therefore, our LL values may not reflect entirely the pathological changes inside the brain. Secondly, the signal to noise ratio of MRSI is proportional to the scanning frequency, meaning our metabolites' quantification is not entirely accurate. Moreover, spectroscopic signal scales can differ from patient to patient, resulting in large metabolite variations. In order to obtain true metabolites concentrations, we would have to measure, for each patient, the T1 and T2 relaxation times for each metabolite, which would be impossible in clinical practice. To overcome some of these limitations, we use as features all three metabolite ratios (NAA/Cho, NAA/Cre, Cho/Cre). By doing so, we expect to retain sufficient valuable information to conduct our analysis.

When comparing classification tasks from a computational point of view, LDA is clearly the winner as the training period last only 3 hours using a computer with 8 threads. Training both SVM-rbf and RF took around 20 days in total and it was done using 60 threads, meaning LDA is approximately 600 times faster than SVM-rbf or RF. Also, the maximum F1-scores for RR vs. PP and RR vs. SP were obtained with LDA and SVM-rbf, suggesting that a linear classifier performs equally good as a non-linear classifier in these cases.

This study is a proof of concept that investigates the added value of MR metabolites combined with clinical data and lesion loads, in classifying MS patients and healthy controls. Clinical data is routinely collected by doctors, lesion load is a known marker of neurodegeneration, while MR metabolites have been shown to provide high specificity of MS pathology. In order to better understand the underlying MS pathological mechanisms, we used three different machine learning methods, one linear and two non-linear, and had a strict quality control for extracting metabolic features. Despite all our efforts, averaging metabolite ratios over "good quality" voxels provides only moderate biomarkers for discriminating MS groups (i.e. RR vs. PP). In general, combining patient age, disease duration, EDSS, and averaged metabolic ratios, leads to the highest classification results. We believe extracting metabolic information from specific brain sub-regions of the MRSI grid (e.g NAWM) should provide a more detailed view of MS

pathology and help the classification tasks. Therefore, further investigations about the MS patients' evolution will be done in the future on sub-regions metabolite quantification, DTI-based brain connectivity metrics, patient treatment, and multi-class classification.

5 CONCLUSIONS

In this paper, we performed nine binary classification tasks and report F1-scores and BAR values after learning linear and non-linear classifiers on combinations of clinical data, lesion loads, and metabolic features. We presented a simple method to compute metabolic features by averaging metabolite ratios over "good quality" voxels of a MRSI grid. Using linear mixed-effects models we found that the MS course is statistically significant in the evolution of four response variables: Lesion Load, NAA/Cre, NAA/Cho, and Cho/Cre ratios. Our results showed that the best classifier for discriminating CIS from RR or RR+SP is SVM-rbf trained on clinical data and lesion loads. We also showed that discriminating RR from PP or SP with high accuracy is possible when training LDA on clinical data. For RR vs. PP, adding metabolic features will not change the results, while for RR vs. SP, adding metabolic features and lesion loads will slightly improve the results.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

FUNDING

This work was funded by European project EU MC ITN TRANSACT 2012 (no. 316679) and the ERC Advanced Grant BIOTENSORS nr.339804. EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

6 REFERENCES

REFERENCES

Arnold, D., De Stefano, N., Narayanan, S., and Matthews, P. (2000). Proton mr spectroscopy in multiple sclerosis. *Neuroimaging clinics of North America* 10, 789–98

Bates, D. M. (2010). lme4: Mixed-effects modeling with r. *URL http://lme4. r-forge. r-project. org/book*

Bitsch, A., Bruhn, H., Vougioukas, V., Stringaris, A., Lassmann, H., Frahm, J., et al. (1999). Inflammatory cns demyelination: histopathologic correlation with in vivo quantitative proton mr spectroscopy. *American Journal of Neuroradiology* 20, 1619–1627

Bjartmar, C., Kinkel, R. P., Kidd, G., Rudick, R. A., and Trapp, B. D. (2001). Axonal loss in normal-appearing white matter in a patient with acute ms. *Neurology* 57, 1248–1252

Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32

Compston, A. and Coles, A. (2008). Multiple sclerosis. *The Lancet* 372, 1502–1518. doi:10.1016/S0140-6736(08)61620-7

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning* 20, 273–297

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods* (Cambridge university press)

Davie, C., Barker, G., Thompson, A., Tofts, P., McDonald, W., and Miller, D. (1997). 1h magnetic resonance spectroscopy of chronic cerebral white matter lesions and normal appearing white matter in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* 63, 736–742

Davie, C., Hawkins, C., Barker, G., Brennan, A., Tofts, P., Miller, D., et al. (1994). Serial proton magnetic resonance spectroscopy in acute multiple sclerosis lesions. *Brain* 117, 49–58

De Stefano, N. and Filippi, M. (2007). Mr spectroscopy in multiple sclerosis. *Journal of Neuroimaging* 17, 31S–35S

Doyle, T. J., Pathak, R., Wolinsky, J. S., and Narayana, P. A. (1995). Automated proton spectroscopic image processing. *Journal of Magnetic Resonance, Series B* 106, 58–63

Filippi, M., Horsfield, M., Morrissey, S., MacManus, D., Rudge, P., McDonald, W., et al. (1994). Quantitative brain mri lesion load predicts the course of clinically isolated syndromes suggestive of multiple sclerosis. *Neurology* 44, 635–635

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 179–188

Fu, L., Matthews, P., De Stefano, N., Worsley, K., Narayanan, S., Francis, G., et al. (1998). Imaging axonal damage of normal-appearing white matter in multiple sclerosis. *Brain* 121, 103–113

Hannoun, S., Bagory, M., Durand-Dubief, F., Ibarrola, D., Comte, J.-C., Confavreux, C., et al. (2012). Correlation of diffusion and metabolic alterations in different clinical forms of multiple sclerosis. *PLoS One* 7, e32525

He, J., Inglese, M., Li, B. S., Babb, J. S., Grossman, R. I., and Gonen, O. (2005). Relapsing-remitting multiple sclerosis: Metabolic abnormality in nonenhancing lesions and normal-appearing white matter at mr imaging: Initial experience 1. *Radiology* 234, 211–217

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical journal* 50, 346–363

Husted, C., Goodin, D., Hugg, J., Maudsley, A. A., Tsuruda, J., De Bie, S., et al. (1994). Biochemical alterations in multiple sclerosis lesions and normal-appearing white matter detected by in vivo 31p and 1h spectroscopic imaging. *Annals of neurology* 36, 157–165

Inglese, M., Li, B. S., Rusinek, H., Babb, J. S., Grossman, R. I., and Gonen, O. (2003). Diffusely elevated cerebral choline and creatine in relapsing-remitting multiple sclerosis. *Magnetic resonance in medicine* 50, 190–195

Jain, S., Sima, D. M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., et al. (2015). Automatic segmentation and volumetry of multiple sclerosis brain lesions from mr images. *NeuroImage: Clinical* 8, 367–375

Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., et al. (2016). Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses. *Frontiers in Neuroscience* 10, 478

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2015). Package lmertest. *R package version* , 2–0

Larsson, H., Christiansen, P., Jensen, M., Frederiksen, J., Heltberg, A., Olesen, J., et al. (1991). Localized in vivo proton spectroscopy in the brain of patients with multiple sclerosis. *Magnetic resonance in medicine* 22, 23–31

Leary, S. M., Davie, C. A., Parker, G. J., Stevenson, V. L., Wang, L., Barker, G. J., et al. (1999). 1h magnetic resonance spectroscopy of normal appearing white matter in primary progressive multiple sclerosis. *Journal of neurology* 246, 1023–1026

Lublin, F. D., Reingold, S. C., et al. (1996). Defining the clinical course of multiple sclerosis results of an international survey. *Neurology* 46, 907–911

McAlpine, D. and Compston, A. (2005). *McAlpine's multiple sclerosis* (Elsevier Health Sciences)

McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H.-P., Lublin, F. D., et al. (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of neurology* 50, 121–127

Miller, D. H., Chard, D. T., and Ciccarelli, O. (2012). Clinically isolated syndromes. *The Lancet Neurology* 11, 157–169

Muthuraman, M., Fleischer, V., Kolber, P., Luessi, F., Zipp, F., and Groppa, S. (2016). Structural brain network characteristics can differentiate cis from early rrms. *Frontiers in neuroscience* 10

Narayana, P. A. (2005). Magnetic resonance spectroscopy in the monitoring of multiple sclerosis. *Journal of Neuroimaging* 15, 46S–57S

Narayana, P. A., Doyle, T. J., Lai, D., and Wolinsky, J. S. (1998). Serial proton magnetic resonance spectroscopic imaging, contrast-enhanced magnetic resonance imaging, and quantitative lesion volumetry in multiple sclerosis. *Annals of neurology* 43, 56–71

Narayana, P. A., Wolinsky, J. S., Rao, S. B., He, R., Mehta, M., et al. (2004). Multicentre proton magnetic resonance spectroscopy imaging of primary progressive multiple sclerosis. *Multiple Sclerosis* 10, S73–S78

Narayanan, S., Fu, L., Pioro, E., De Stefano, N., Collins, D., Francis, G., et al. (1997). Imaging of axonal damage in multiple sclerosis: spatial distribution of magnetic resonance imaging lesions. *Annals of neurology* 41, 385–391

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria. *Annals of neurology* 69, 292–302

Polman, C. H., Reingold, S. C., Edan, G., Filippi, M., Hartung, H.-P., Kappos, L., et al. (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the mcdonald criteria. *Annals of neurology* 58, 840–846

Poullet, J.-B. (2008). Quantification and classification of magnetic resonance spectroscopic data for brain tumor diagnosis. *Katholic University of Leuven*

Poullet, J.-B., Sima, D., Luts, J., Garcia, M. O., Croitor, A., and Van Huffel, S. (2008). Manual: Simulation package based on in vitro databases (spid)

Poullet, J.-B., Sima, D. M., Simonetti, A. W., De Neuter, B., Vanhamme, L., Lemmerling, P., et al. (2007). An automated quantitation of short echo time mrs spectra in an open source software environment: Aqses. *NMR in Biomedicine* 20, 493–504

Rovira, À., Auger, C., and Alonso, J. (2013). Magnetic resonance monitoring of lesion evolution in multiple sclerosis. *Therapeutic advances in neurological disorders* 6, 298–310

Sajja, B. R., Wolinsky, J. S., and Narayana, P. A. (2009). Proton magnetic resonance spectroscopy in multiple sclerosis. *Neuroimaging clinics of North America* 19, 45–58

Sarchielli, P., Presciutti, O., Pelliccioli, G., Tarducci, R., Gobbi, G., Chiarini, P., et al. (1999). Absolute quantification of brain metabolites by proton magnetic resonance spectroscopy in normal-appearing white matter of multiple sclerosis patients. *Brain* 122, 513–521

Sava, C., Anca, R., Sima, D. M., Poullet, J.-B., Wright, A. J., Heerschap, A., et al. (2011). Exploiting spatial information to estimate metabolite levels in two-dimensional mrsi of heterogeneous brain lesions. *NMR in Biomedicine* 24, 824–835

Scalfari, A., Neuhaus, A., Degenhardt, A., Rice, G. P., Muraro, P. A., Daumer, M., et al. (2010). The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability. *Brain* 133, 1914–1929

Sicotte, N. L., Voskuhl, R. R., Bouvier, S., Klutch, R., Cohen, M. S., and Mazziotta, J. C. (2003). Comparison of multiple sclerosis lesions at 1.5 and 3.0 tesla. *Investigative radiology* 38, 423–427

Suhy, J., Rooney, W., Goodkin, D., Capizzano, A., Soher, B., Maudsley, A. A., et al. (2000). 1h mrsi comparison of white matter and lesions in primary progressive and relapsing-remitting ms. *Multiple sclerosis* 6, 148–155

Sundin, T., Vanhamme, L., Van Hecke, P., Dologlou, I., and Van Huffel, S. (1999). Accurate quantification of 1 h spectra: From finite impulse response filter design for solvent suppression to parameter estimation. *Journal of Magnetic Resonance* 139, 189–204

Tartaglia, M., Narayanan, S., De Stefano, N., Arnaoutelis, R., Antel, S., Francis, S., et al. (2002). Choline is increased in pre-lesional normal appearing white matter in multiple sclerosis. *Journal of neurology* 249, 1382–1390

Tiberio, M., Chard, D., Altmann, D., Davies, G., Griffin, C., McLean, M., et al. (2006). Metabolite changes in early relapsing–remitting multiple sclerosis. *Journal of neurology* 253, 224–230

Tourbah, A., Stievenart, J.-L., Abanou, A., Iba-Zizen, M.-T., Hamard, H., Lyon-Caen, O., et al. (1999). Normal-appearing white matter in optic neuritis and multiple sclerosis: a comparative proton spectroscopy study. *Neuroradiology* 41, 738–743

Wattjes, M., Harzheim, M., Lutterbey, G., Klotz, L., Schild, H., and Träber, F. (2007). Axonal damage but no increased glial cell activity in the normal-appearing white matter of patients with clinically isolated syndromes suggestive of multiple sclerosis using high-field magnetic resonance spectroscopy. *American Journal of Neuroradiology* 28, 1517–1522

Wattjes, M. P., Harzheim, M., Lutterbey, G. G., Bogdanow, M., Schild, H. H., and Träber, F. (2008). High field mr imaging and 1h-mr spectroscopy in clinically isolated syndromes suggestive of multiple sclerosis. *Journal of neurology* 255, 56–63

Wolinsky, J. S., Narayana, P. A., and Fenstermacher, M. J. (1990). Proton magnetic resonance spectroscopy in multiple sclerosis. *Neurology* 40, 1764–1764

7 APPENDIX

	NAA/Cho			NAA/Cre			Cho/Cre			All 3 metabolites		
	BAR	SPE	SEN	BAR	SPE	SEN	BAR	SPE	SEN	BAR	SPE	SEN
HC vs. CIS	47	0	94	46	15	78	61	39	83	53	39	67
HC vs. RR	50	94	6	55	82	28	50	100	0	52	76	28
HC vs. PP	76	80	72	78	72	83	45	29	61	77	82	72
HC vs. RR + SP	52	98	6	60	92	28	50	100	0	59	90	28
HC vs. RR + PP	61	89	33	66	88	44	50	100	0	52	88	16
CIS vs. RR	52	95	10	50	100	0	50	99	0	52	88	16
CIS vs. RR + SP	51	100	2	49	99	0	50	100	0	54	94	15
RR vs. PP	59	37	81	63	38	88	48	2	95	63	49	77
RR vs. SP	57	53	62	65	62	69	39	0	79	66	62	70

Table 6. Balanced accurary rates (BAR), sensitivity (SEN), and specificity (SPE) values, for all 9 classification tasks (rows) after training LDA using only metabolic ratios. Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, values between 85 and 89 are coloured in dark gray, while values higher than 90 are coloured in very dark gray.

	CIS vs. RR			CIS vs. RR+SP			RR vs. PP			RR vs. SP		
	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF
M	52	68	49	54	63	49	63	28	59	66	66	63
LL	48	70	52	50	73	56	43	12	58	74	75	68
Age + DD	66	75	68	66	83	70	67	38	62	75	76	71
Age + DD + EDSS	71	80	67	77	89	69	81	78	70	84	85	84
Age + DD + EDSS + LL	79	85	73	81	92	76	71	72	69	86	86	85
Age + DD + EDSS + M	72	76	66	81	82	70	80	81	71	86	87	84
Age + DD + EDSS + LL + M	78	80	71	82	83	73	78	78	68	86	86	86

Table 7. BAR values for classification tasks involving only MS patients (columns). Abbreviations: M = all three average metabolic ratios; Age = patient age; DD = disease duration; LL = lesion load; EDSS = Expanded Disability Status Scale.Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, values between 85 and 89 are coloured in dark gray, while values higher than or equal to 90 are coloured in very dark gray.

	CIS vs. RR			CIS vs. RR+SP			RR vs. PP			RR vs. SP		
	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF
M	16	79	8	15	67	0	77	56	78	70	75	72
LL	0	80	30	0	80	23	87	16	78	77	80	67
Age + DD	41	77	49	36	84	46	84	75	78	74	70	70
Age + DD + EDSS	56	82	44	62	92	43	84	75	80	80	83	81
Age + DD + EDSS + LL	69	87	56	69	93	57	81	69	83	85	84	85
Age + DD + EDSS + M	59	74	41	74	79	44	84	76	82	84	85	84
Age + DD + EDSS + LL + M	67	79	49	72	77	51	83	75	81	87	87	86

Table 8. Sensitivity values for classification tasks involving only MS patients (columns). Abbreviations: M = all three average metabolic ratios; Age = patient age; DD = disease duration; LL = lesion load; EDSS = Expanded Disability Status Scale. Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, values between 85 and 89 are coloured in dark gray, while values higher than or equal to 90 are coloured in very dark gray.

	CIS vs. RR			CIS vs. RR+SP			RR vs. PP			RR vs. SP		
	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF	LDA	SVM-rbf	RF
M	88	57	89	94	59	98	49	0	40	62	56	54
LL	96	60	75	100	66	89	0	7	37	70	70	69
Age + DD	91	74	87	96	83	94	50	0	46	76	82	72
Age + DD + EDSS	87	79	89	91	87	95	78	81	60	89	87	86
Age + DD + EDSS + LL	89	83	90	92	90	95	60	75	55	87	87	85
Age + DD + EDSS + M	85	78	91	89	86	95	75	86	60	88	88	84
Age + DD + EDSS + LL + M	88	81	93	92	89	96	74	82	56	85	86	85

Table 9. Specificity values for classification tasks involving only MS patients (columns). Abbreviations: M = all three average metabolic ratios; Age = patient age; DD = disease duration; LL = lesion load; EDSS = Expanded Disability Status Scale. Values between 75 and 79 are coloured in light gray, values between 80 and 84 are coloured in medium gray, values between 85 and 89 are coloured in dark gray, while values higher than or equal to 90 are coloured in very dark gray.

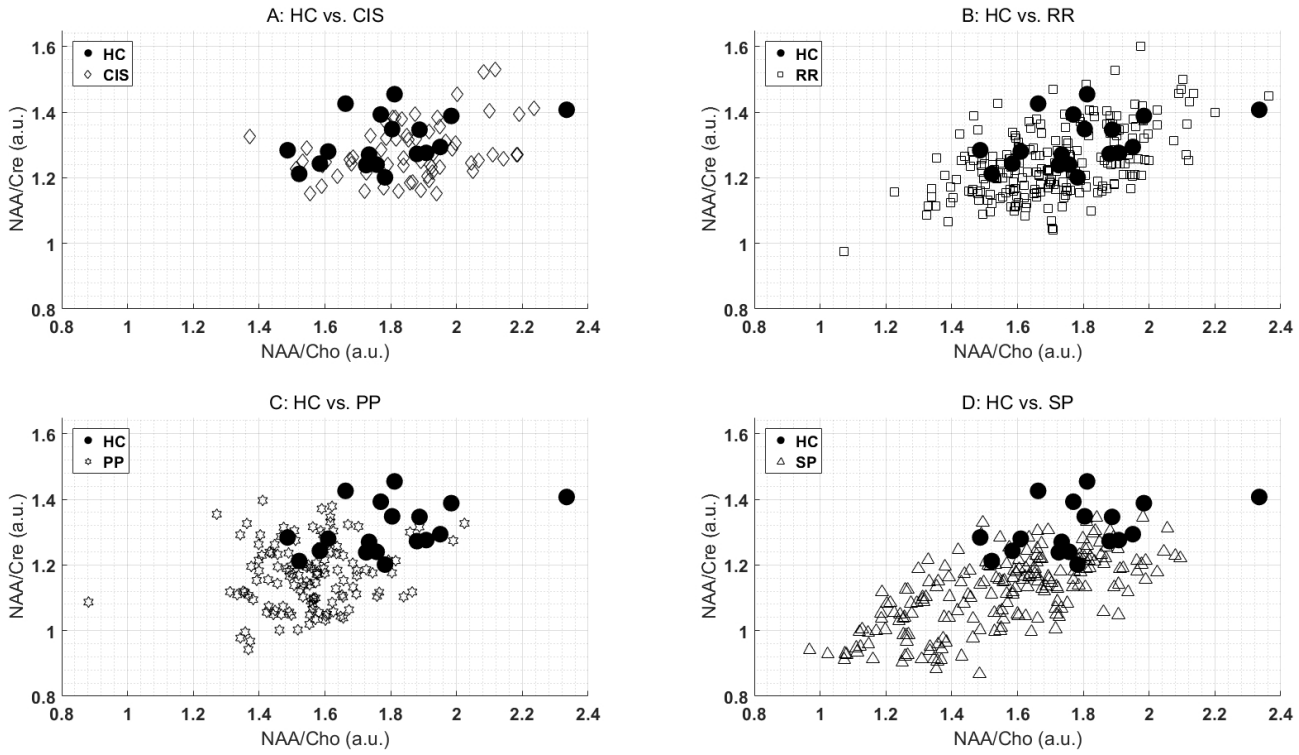

Figure 2. HC vs. MS groups in 2-D feature space: x-axis is NAA/Cho and y-axis is NAA/Cre.

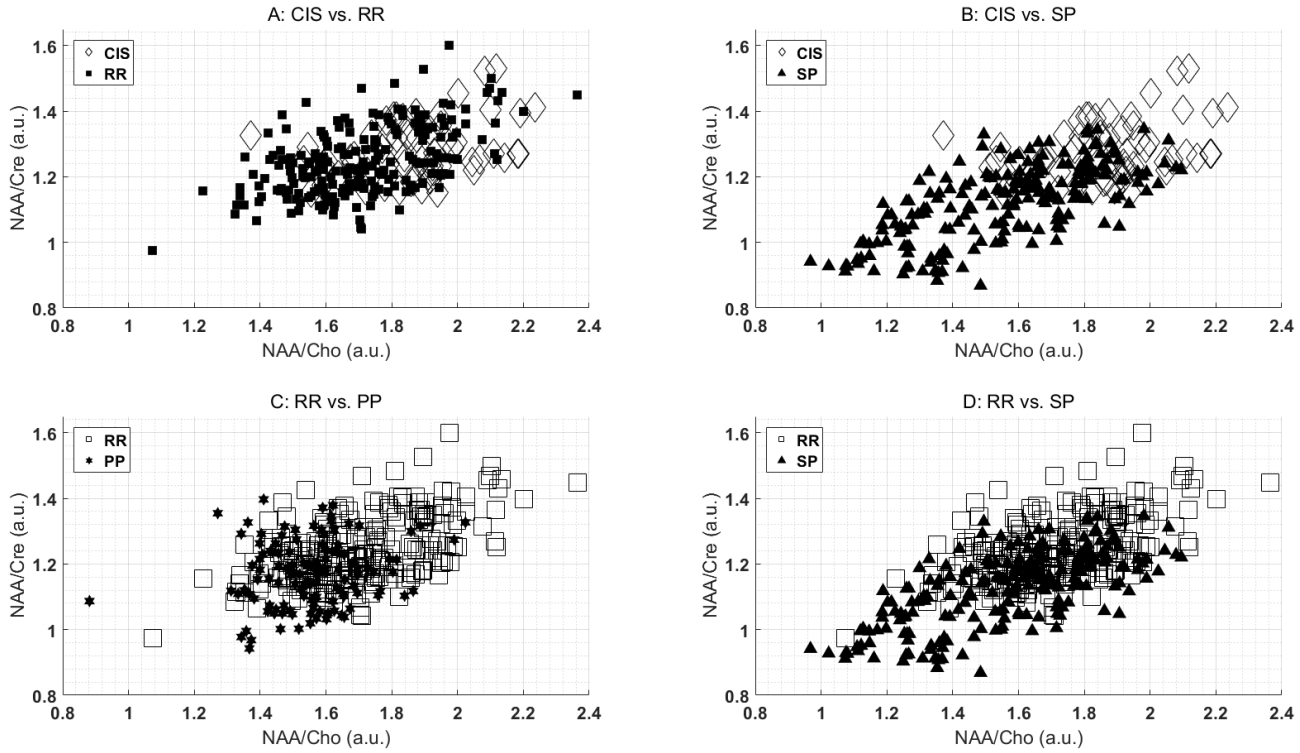

Figure 3. Comparison of MS groups in 2-D feature space: x-axis is NAA/Cho and y-axis is NAA/Cre.

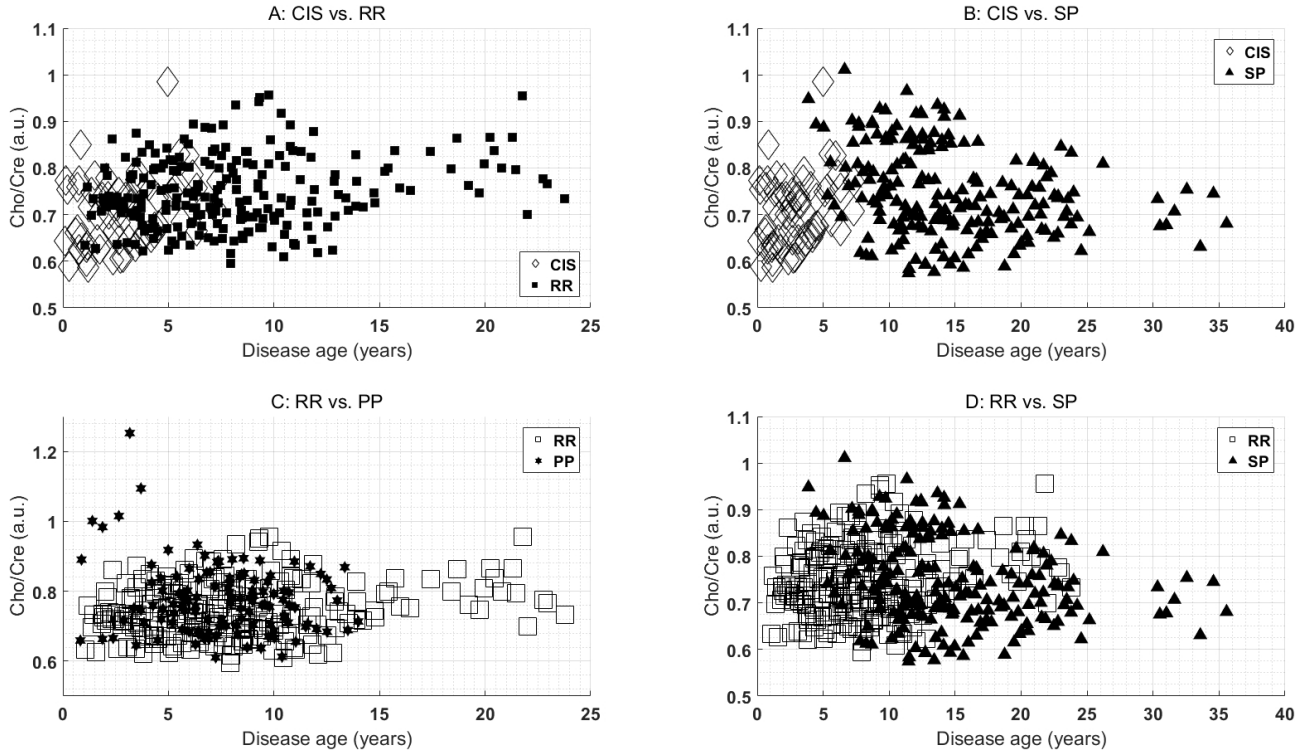

Figure 4. Comparison of MS groups in 2-D feature space: x-axis is disease age and y-axis is Cho/Cre.

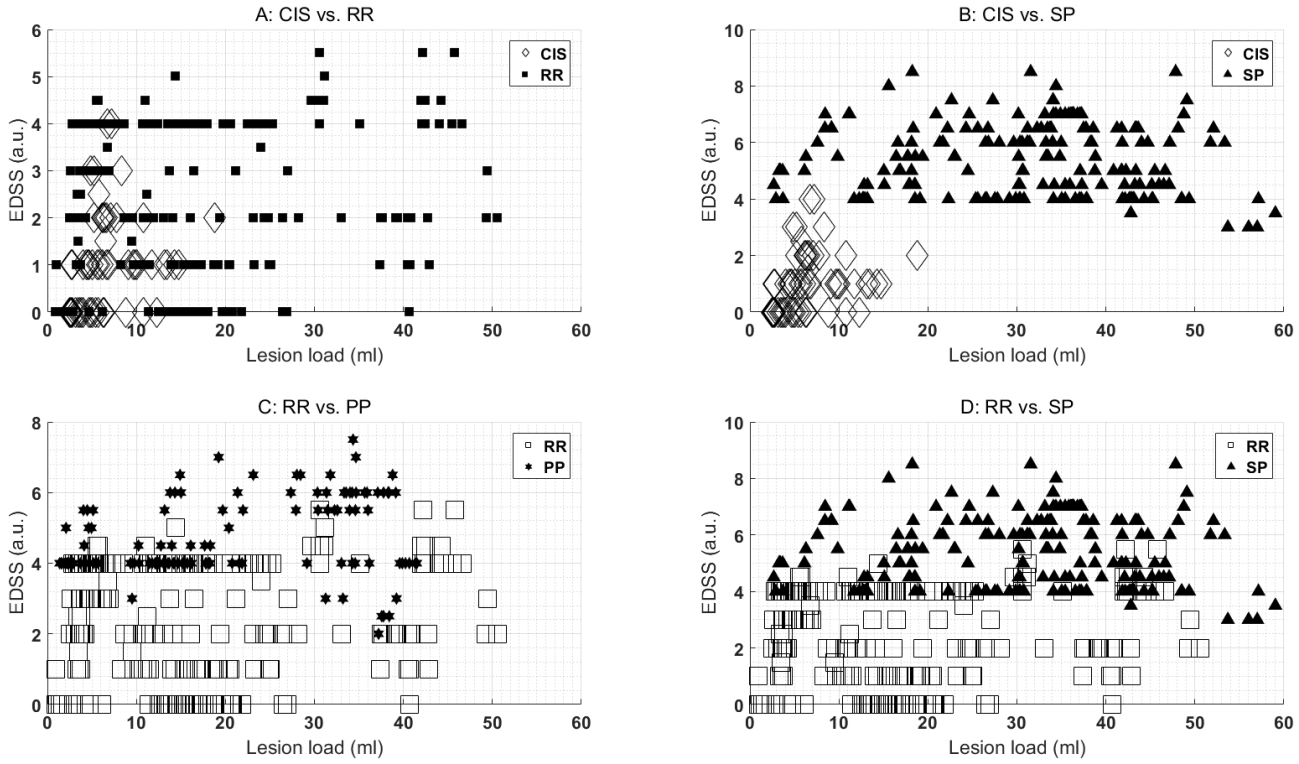

Figure 5. Comparison of MS groups in 2-D feature space: x-axis is lesion load and y-axis is EDSS.